Psychiatry Research 175 (2010) 266–270

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Sex-based differences in answering strategy and the influence of cross-sex hormones

Muirne Caitlin Shonagh Paap [a,b,*], Ira Ronit Haraldsen [b]

[a] Institute of Psychiatry, University of Oslo, Rikshospitalet, Oslo, Norway
[b] GID Clinic, Department of Neuropsychiatry and Psychosomatic Medicine, Rikshospitalet, Oslo, Norway

ARTICLE INFO

Article history:
Received 19 November 2008
Received in revised form 10 June 2009
Accepted 29 July 2009

Keywords:
Sex differences
Cognition
Guessing
Gender identity
Mathematical ability
Risk-taking

ABSTRACT

We investigated whether sex differences in answering strategy occur in normal controls (C). Furthermore, it was tested whether these sex differences were subject to change over time, and whether they were associated with hormonal treatment at time points 2 and 3 in patients with Gender Identity Disorder (GID). Two subtests measuring arithmetic ability were used: arithmetic aptitude (AA) and arithmetic operations (AO). Both the controls ($n=29$) and GID patients ($n=33$) were tested at baseline (T1), three months (T2) and 12 months (T3) after the start of hormonal treatment in the GID group. A repeated measures analysis of variance showed no differences between C males and females, for T1 and T2. At T3, C males guessed more than C females. At baseline, GID males and C males left an equal number of items unanswered. However, when being retested, C males left fewer items unanswered than GID males. No difference was found between C females and GID females at any time point. Our results suggest that healthy adult males become more confident when they are retested, and seem to adjust their answering strategy accordingly. Moreover, hormonal treatment of healthy adult GID patients born male is associated with a lack of adjustment in answering strategy.

1. Introduction

Sex differences in cognitive abilities have been a 'hot topic' for decades. The finding that boys perform better than girls on tests of mathematic ability (Maccoby and Jacklin, 1974) has received much of attention from researchers in different fields since its publication in 1974 and seems to be a persistent finding, though researchers have stressed in recent years that the differences are only moderate to small (Hyde, 2005; Torres et al., 2006; Liu et al., 2008). Several approaches have been employed in the attempt to explain sex differences in cognitive abilities: a focus on test characteristics, situational aspects, personality traits and hormonal influences (Kimball, 1989; Ben-Shakhar and Sinai, 1991; Meurling et al., 2000; Liu et al., 2008).

In attempts to explain why sex differences occur, the question arises whether there are 'true' differences between the sexes or whether these differences are artefacts of the test. One way of examining such test artefacts, or test bias, is to look at each item individually. If a person from one sex is more likely to answer the item correctly than a person of the other sex who has the same mathematical ability, the item is biased. This item bias is referred to as 'Differential Item Functioning' (DIF) or 'Differential Item Performance' (DIP) (Doolittle and Cleary, 1987; Embretson and Reise, 2000; Gierl et al., 2003). DIF analyses can help identify biased items that could be subsequently removed, thus reducing sex-related artefacts. When it comes to the situational aspect, it has been consistently found that boys tend to perform better than girls on standardized math tests (such as national examinations or assessments) whereas girls perform better than boys on tests conducted in a classroom setting (Kimball, 1989; Ben-Shakhar and Sinai, 1991; McGillicuddy-De Lisi and De Lisi, 2002; Liu et al., 2008). At least a partial explanation might be offered by the 'novelty versus familiarity hypothesis' (Kimball, 1989): males would perceive novel situations as a challenge, whereas females would feel insecure. This finding might be related to differences in personality traits, such as aggression and risk-taking, which seem to differ between the sexes (see Ben-Shakhar and Sinai, 1991; Hyde, 2005), and could influence test performance through the willingness to guess (take a risk), which boys are more inclined to do than girls (Hudson, 1986; Ben-Shakhar and Sinai, 1991).

In healthy males and females, sex differences in cognitive performance have been well documented (Linn and Petersen, 1985; Silverman and Phillips, 1993; McKeever, 1995; Voyer et al., 1995; James and Kimura, 1997). Studies linking sex hormones to differences in cognition between the sexes have made a distinction between so-called 'organizing effects' (influence on embryonic brain development) and 'activating effects' (modulation of cognitive performance in adult life); but a unifying biological model has yet to be presented (Geschwind and Galaburda, 1985; Christiansen and Knussmann, 1987; Hampson and Kimura, 1988; Gouchie and Kimura, 1990; Williams and Meck, 1990; Sherwin, 1994; Meurling et al., 2000). One way of

* Corresponding author. Clinic for Gender Identity Disorder, Department of Neuropsychiatry and Psychosomatic Medicine, Rikshospitalet, NO-0027 Oslo, Norway. Tel.: +47 23074160; fax: +47 23074170.
E-mail address: muirne@nxdomain.nl (M.C.S. Paap).

0165-1781/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.psychres.2009.07.020

discerning between the influences of organizing and activating effects, is to manipulate the level of sex hormones in an adult person. Studies using this method focus on patients with congenital hormone disorders, such as Congenital Adrenal Hyperplasia (Cem Demirci, 2008) as well as physically healthy persons who seek cross-sex hormone treatment; these are mostly patients with Gender Identity Disorder (GID). GID is characterized by a discrepancy between biological sex and gender identification, in spite of hormonal levels that are normal with respect to the biological sex. Both studies on congenital hormone disorders and GID have produced contradictory results when it comes to the organizing/activating effects of sex hormones on cognitive functioning (Resnick et al., 1986; Van Goozen et al., 1995; Slabbekoorn et al., 1999; Malouf et al., 2006; Puts et al., 2008), although the most recent studies of GID patients have consistently found that GID patients showed a pattern of cognitive performance similar to their biological sex, in spite of current hormonal treatment (Van Goozen et al., 2002; Haraldsen et al., 2003; Haraldsen et al., 2005).

In this study, which is based on data collected for a previous study (Haraldsen et al., 2005), we are interested in possible differences in answering strategies between the sexes on math tests conducted in an unfamiliar setting. Our aim is to test three hypotheses:

1. We expect males to leave fewer items unanswered than females.
2. We hypothesize that guessing strategy is stable over time.
3. We expect GID patients treated with cross-sex hormones to employ a guessing strategy that is similar to the one used by controls of the same biological sex.

2. Method

2.1. Subjects

This study included 62 participants. The control (C) group members ($n=29$, 15 females, 14 males) were healthy, heterosexual Norwegian participants with no lifetime diagnosis of gender identity disorder (GID). They were either high school graduates, military recruits from the armed forces, college students or employees of the University of Oslo. They were recruited by advertisement. The patient group consisted of somatically healthy individuals diagnosed with GID who consecutively sought sex reassignment surgery (SRS) in Norway from 1996 to 1998 ($n=33$, 21 females, 12 males). All included GID patients were diagnosed by two psychiatrists, fulfilling criteria A to D in DSM-IV-TR (American Psychiatric Association, 2000) from childhood onwards.

All participants were chromosomally and endocrinologically screened, and medication-free. No GID patient had ever received previous cross-sex hormone treatment. Participants with any endocrinological, genetic, neurological or major psychiatric comorbidity were excluded [$n=3$, (2 delusional disorders, 1 XXY anomalies)]. All participants were Caucasians.

The median age of GID males ($M=29$, IQR = 8.5) differed from that of C males ($M=20$, IQR = 3.0), Mann–Whitney $U[36]=35$, $P<0.001$. The median age of GID females ($M=23$, IQR = 7.5) did not differ significantly from the median age of the C females ($M=21$, IQR = 26.0), Mann–Whitney $U[45]=176.5$, $P=0.242$. All participants had completed high school. Average number of correct answers was used as a proxy for performance level. The control group showed a higher level than the GID group; mean difference = 1.57, $t=3.78$, $P<0.001$.

2.2. Hormonal treatment

All male-born GID patients ($n=12$) received 50 µg of oral ethinylestradiol (Etifollin) daily during the first 3 months of treatment, and thereafter 100 µg daily. All female-born GID patients ($n=21$) received 180 mg testosterone enantate (Primoteston-Depot) as an intramuscular injection every third week.

2.3. Neuropsychological testing: arithmetic aptitude and arithmetic operations

The current study is based on data collected for a previous study (Haraldsen et al., 2005). A full description of all cognitive tests used in the neuropsychological testing battery can be found there. The two subtests used in this study stem from the factor 'reasoning, general,' which is included in the officially distributed "Kit of Factor-Referenced Cognitive Tests" by ETS [Educational Testing Service (www.ets.org), (Ekstrom et al., 1976)]. The factor is based on three subtests, of which we used 'arithmetic aptitude' (AA) and 'arithmetic operations' (AO), each consisting of 15 items. We chose not to use 'mathematics aptitude' since it shows a considerable overlap in difficulty with 'arithmetic aptitude'. 10 min are available for each subtest. In the first subtest (AA) the subject has to calculate the answer and select it from five alternative answers. In AO, the subject picks the correct arithmetic operation required for a given result (e.g. addition, subtraction).

All participants were tested on three occasions: baseline (T1), 3 months (T2), and 12 months (T3), respectively, after the GID patients had started with hormone treatment. All C females were tested during the first 2 weeks of their menstrual cycle. Each test session started at 0900 h and lasted for 3 h with two 15-min breaks after the first and second hour. The order of test presentation was random and was administered by one of two trained test assistants.

2.4. Missing data

Missing data occurred in this study: 2 out of 12 GID males, 7 out of 14 C males, 7 out of 21 GID females and 3 out of 15 C females did not complete 'arithmetic' on at least one measurement occasion. This is considered missing at random; each participant showed up at each measurement occasion, but it frequently occurred that not all tests were completed by all participants. This could be explained by lack of motivation or fatigue. Analyses revealed that no test was more likely to be left unanswered than any other test. Moreover, no significant difference was found in the mean number of unanswered items at baseline, when comparing the participants who had completed all three measurements ($n=43$) with the participants who had only completed the first two ($n=19$), by means of a t-test ($t=-0.43$, $P=0.67$).

2.5. Statistics

To investigate possible differences in guessing tendency between C males and C females, and to see whether hormonal treatment would have an impact on guessing tendency, a repeated measures analysis of variance (ANOVA, GLM repeated measures, SPSS 15.0, 2007) was applied. In this model, the number of unanswered items (averaged over the two subtests) was the dependent variable, and time (baseline, 3 months, 12 months) served as the within-subject factor. One item from the AA subtest was deleted before proceeding to the analyses, since it was answered incorrectly by almost all subjects. We corrected this by multiplying the total number of unanswered items by 15/14 for the AA subtest. Between-subject factors were biological sex (male, female) and group (GID, C). Age and the average number of correct answers were included as covariates. For 'time' (baseline, 3 months, 12 months), the 'repeated' contrast was used. Planned comparisons were used (Stevens, 2002) to investigate whether C males guessed more than C females, and whether GID patients undergoing hormonal treatment showed a pattern comparable to the controls with regard to their biological sex.

The reported significance values for the repeated measures ANOVAs in this study were based on the Huynh–Feldt estimator of epsilon (Huynh and Feldt, 1976), which is to be preferred to the more conservative Greenhouse-Geisser estimator when the estimated epsilon is above 0.70 (Stevens, 2002). To correct for capitalization on chance, the Bonferroni–Holm procedure was used (Shaffer, 1995), which is less conservative and thus more powerful than the simple Bonferroni procedure and which can always be used instead of the classical Bonferroni procedure (Shaffer, 1995; Ekenstierna, 2004). An alpha of 0.05 was used.

3. Results

3.1. Hormonal levels

In Table 1 the relevant endocrinological data of controls and GID patients are summarized. No significant differences were found between the GID patients and their sex-matched controls before treatment. The values were within the range of laboratory standard values. For the GID males, the biological effect of the estrogen treatment was clearly demonstrated by the large, and significant, increase in serum sex hormone binding globulin (SHBG) levels and a drop in testosterone levels. For the GID females, testosterone treatment led to a significant increase in testosterone concentrations from normal female to normal male levels, and this was accompanied by an expected drop in SHBG levels. The serum levels of estrogen (E2) did not show any changes for either of the sexes. That is likely connected to the fact that the assay does not detect changes in ethinylestradiol. For more detailed information about the laboratory methods, see the study by Haraldsen et al. (2005).

3.2. Description of the data

It can be seen in Table 2 that C males showed marked changes over time: the number of unanswered items decreased, while both the number of correct and incorrect responses increased. In contrast, C females showed only slight changes in any of the response categories.

In fact, the GID females showed a similar pattern to the C males, though of a lesser magnitude. The most marked pattern is found for the GID males: first they showed an increase in number of unanswered items, and then a decrease that brought them back to baseline level.

3.3. Repeated measures: hypothesis testing

Table 3 shows the results of the final model of the repeated measures analysis. Since the assumption of sphericity was violated, we chose to use the Huyn–Feldt adjustment for epsilon. Factors of interest were the main effects of time (Is there a change over time overall?) and sex (Is there an overall effect of sex, averaged over the time points?); the interaction between time and sex (Is the change over time different for the two sexes?); and especially the three-way interaction between time, sex and group (Do GID patients taking hormones show similar differences between the sexes as controls?).

Our final model included both sex and group as independent variables. It did not include age, since this variable was neither found to be significant when it was controlled for, nor to change any of the effects of interest. The latter was also true for the other covariate, average of unanswered items, and therefore this variable was not included in the final model either. Table 3 contains the results of the analysis according to the earlier stated hypotheses. A significant main effect of time, a significant two-way interaction between sex and group, and a significant three-way interaction between time, sex and group were found. The main effect of time (see Table 3) indicates an overall change in mean number of unanswered items over the three measurement points (the means for T1, T2 and T3 were 7.2, 6.6 and 6.0, respectively). The interaction between sex and group (see Table 3) was caused by the C males and females differing more from each other than the GID males and females (7.1 and 5.1 versus 6.6 and 7.6). The three-way interaction between time, sex and group indicates that the two-way interaction changes over time (see Table 3 and Fig. 1).

The next step was to shed light on the three-way interaction using planned comparisons, in order to find out, for each time point, whether C males and C females differed in their answering strategy (the number of items they left unanswered), and whether GID patients applied a strategy comparable to controls of the same biological sex. At T1, no differences were found. However, at T2, C males guessed more than GID males (mean difference = 2.98, $P = 0.003$, $\alpha = 0.007$) and at T3, C males guessed more than C females (mean difference = 3.96, $P = 0.001$, $\alpha = 0.006$) as well as GID males (mean difference = 4.48, $P < 0.001$, $\alpha = 0.006$). Fig. 1 depicts the three-way interaction graphically: the line for the C males shows the steepest slope, whereas the C female line almost has a slope of 0. Like the C males, the GID females showed a decrease, but their slope is less steep. They did not differ significantly from the C females. The GID males showed a pattern, which was strikingly different from that of the C males, where they initially showed an increase and then a decrease, which brought them back to baseline level.

Table 1
Median (IQR) hormonal levels of the controls, and the GID patients for the three measurement points.

	Controls	GID patients		
		T1	T2	T3
Males				
Testosterone (nM/l)	18.7 (7.2)	20.8 (8.9)	1.1 (17.9)	2.7 (17.3)*
E2 (nM/l)	0.1 (0.03)	0.1 (0.05)	0.08 (0.03)	0.05 (0.05)
SHBG (nM/l)	24.0 (9.0)	25.5 (13.5)	196.0 (91.0)*	214.0 (167.5)*
LH (IU/l)	4.4 (2.4)	4.8 (2.7)	1.1 (4.0)	1.8 (5.0)
FSH (IU/l)	3.6 (2.4)	3.1 (3.4)	1.0 (0)	1.0 (0.9)*
Females				
Testosterone (nM/l)	1.2 (1.4)	1.7 (0.6)	21.9 (18.0)*	29.3 (8.0)*
E2 (nM/l)	0.3 (0.4)	0.4 (0.5)	0.2 (0.08)	0.2 (0.1)*
SHBG (nM/l)	70.0 (51.0)	52.0 (47.0)	25.5 (17.0)*	22.5 (8.0)*
LH (IU/l)	7.9 (9.6)	7.8 (17.6)	4.0 (5.2)*	4.2 (8.6)*
FSH (IU/l)	4.3 (6.3)	5.3 (3.2)	5.8 (3.3)	4.8 (4.7)

* Significant change with respect to T1, based on the Wilcoxon Signed Ranks Test.

Table 2
Raw mean scores (S.D.) and *n* of 'arithmetic' averaged over the two subtests per type of answer by group, biological sex and measurement point.

	T1	*n*	T2	*n*	T3	*n*
Unanswered items						
C males	8.0 (1.66)	14	5.8 (1.95)	10	3.8 (2.39)	8
C females	8.1 (1.85)	15	7.5 (2.57)	13	7.7 (1.90)	12
GID males	7.8 (3.02)	12	8.8 (1.88)	10	8.2 (1.97)	10
GID females	7.7 (3.40)	21	6.1 (3.18)	19	6.5 (3.30)	16
Correct items						
C males	5.0 (1.94)	14	6.3 (1.80)	10	8.1 (1.43)	8
C females	5.0 (1.53)	15	5.0 (1.63)	13	5.1 (1.45)	12
GID males	3.6 (1.50)	12	3.5 (1.06)	10	4.0 (1.40)	10
GID females	3.8 (2.34)	21	4.0 (1.82)	19	4.5 (2.09)	16
Incorrect items						
C males	2.0 (1.66)	14	2.9 (1.76)	10	3.2 (2.41)	8
C females	2.0 (1.55)	15	2.5 (2.17)	13	2.2 (1.66)	12
GID males	3.6 (3.38)	12	2.7 (1.29)	10	2.7 (1.84)	10
GID females	3.5 (2.93)	21	4.9 (3.34)	19	4.0 (3.70)	16

4. Discussion

Existing research directed at explaining differences between males and females in cognitive ability, such as math, could largely be divided into two groups: studies that focus on testing psychology and those that focus on hormonal influences. The former have shown that it is important to carefully scrutinize the test that is used for potential gender bias, to take into account the situational aspect (standardized test versus classroom) and differences in personality traits that could be correlated with the readiness to guess at multiple choice tests (Kimball, 1989; Ben-Shakhar and Sinai, 1991; Liu et al., 2008). In these studies, however, it has not been examined whether these aspects or the influences thereof change when the participant is retested. Retesting is exactly what studies that focus on hormonal influences have paid attention to; the most recent studies in this field have shown that cognitive abilities are not likely to change as an effect of cross-sex hormone treatment (Van Goozen et al., 2002; Haraldsen et al., 2003; Haraldsen et al., 2005). However, in these studies little attention was paid to the potential impact of situational aspects or sex-related differences in personality traits that might influence the test results.

In this study, we set out to help fill a gap in the existing literature, by combining aspects of the testing psychology-focused studies (situational aspect, personality traits) and the hormone-focused studies (retesting, influence of cross-sex hormone treatment). We focused on guessing strategy, which we believed (a) to be related to the sex-specific trait risk taking and (b) depend on levels of self-

Table 3
Results of repeated measures ANOVA for number of unanswered items, $n = 44$.

Effect	*F*-value	*Df**	*P*-value*
Time	3.97	2, 82	0.03**
Sex	0.70	1, 41	0.42
Group	2.80	1, 41	0.10
Time × group	1.33	2, 82	0.27
Time × group	1.96	2, 82	0.16
Sex × group	6.12	1, 41	0.02**
Time × sex × group	6.07	2, 82	0.01**

Note: the results in this table are based on our final model (within-subject factor: time, between-subject factors: group, sex); only effects of interest are listed.

* The Huynh–Feldt adjustment was used; ε_{HF} (time) = 0.84, ε_{HF} (measure) = 1, ε_{HF} (measure * time) = 1.

** Statistically significant.

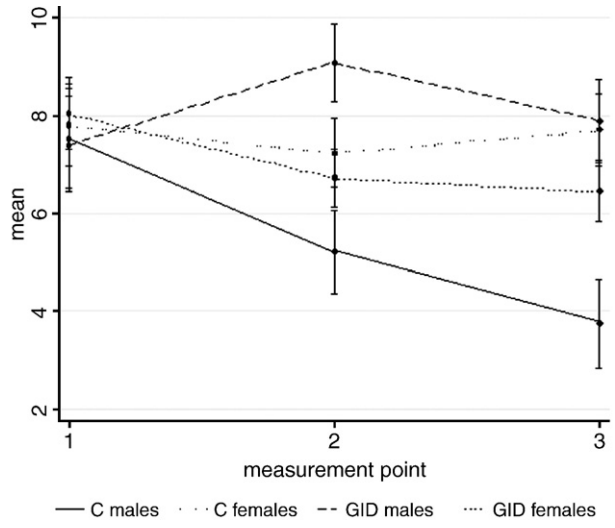

Fig. 1. Mean number unanswered items (based on the GLM repeated measures) ± S.E., by group and measurement point (1 = baseline, 2 = 3 months after start of treatment, 3 = 12 months after start of treatment).

confidence. We expected males to be more confident than females, since the test was both math-related and novel to the participants. Firstly, then, we wanted to see whether we could replicate the finding that men would guess more readily than women. More importantly, we analyzed whether this effect was stable over time. Finally, we investigated whether GID patients treated with cross-sex hormones resembled the controls with regard to guessing strategy and the change herein, if appropriate.

All three hypotheses were rejected. Firstly, there were no significant differences at baseline between the control (C) females and C males. Secondly, C males adjusted their guessing strategy when being retested; at the third measurement, C males were leaving significantly fewer items unanswered than females. Thirdly, GID males undergoing hormonal therapy did not show a pattern comparable to that of the C males. They left significantly more items unanswered than their C counterparts at T2 as well as T3. GID females did not show a pattern that differed from C females.

Our finding that there were no differences in number of items left unanswered at baseline is in conflict with previous research findings that showed differences in answering strategies between males and females on math tests (Ben-Shakhar and Sinai, 1991; McGillicuddy-De Lisi and De Lisi, 2002) as well as with the 'novelty versus familiarity hypothesis' (Kimball, 1989). However, those studies were not conducted in Scandinavia. It has been proposed that gender differences in general might be smaller in the Scandinavian countries, men and women being more equal than in other societies (see Eriksson and Lindholm, 2007), which would translate into weaker stereotypes about sex differences in mathematical ability (Brandell et al., 2005). As a consequence, Scandinavian women might not be subjected to stereotype threat as frequently as women from other countries (Inzlicht and Ben-Zeev, 2000; Keller and Dauenheimer, 2003). Indeed, in a Swedish study by Wester and Henriksson (2000), no support was found for the idea that males are more inclined to guess than females. The lack of differences at baseline in our study might be interpreted as an influence of culture on feelings of self-confidence. It might be argued, that 'even' males need a confidence boost when a testing situation is very new to them (e.g. first measurement). Interestingly, our findings do indicate that Norwegian males become more confident when being retested.

The GID males showed a sex-atypical pattern: they left more items unanswered than C males at both T2 and T3. Their conservative answering strategy might be explained by the estrogen treatment impacting the level of confidence or daringness in these patients.

There were some participants who, although they did show up for the test session, did not complete the 'arithmetic tests'. We feel confident, however, that our results were not biased as an effect of this because we found that the participants were not more likely to skip the arithmetic tests than any other tests in the testing battery. Moreover, further scrutiny of the data revealed that there was no significant difference between the mean number of unanswered items of the participants who had completed all three measurements and the mean number of the participants who had only completed the first two tests.

The present study was not designed to match the controls and patients for age and performance level; however, neither the effect of age nor that of performance level was found to be significant when controlled for, nor did they confound the effects of interest. Neither does the nature of our design permit us to draw firm conclusions as to whether it is the condition (GID) itself or the hormonal treatment that caused the observed differences between GID and C males. A study whereby the control group would consist of GID patients not yet receiving treatment might further elucidate this issue. Nonetheless, our results are quite surprising with regard to the latest research findings on cognitive performance in GID patients, which interpreted the finding that cross-sex hormone treatment is not accompanied by a change in overall cognitive performance in GID patients as evidence for a resemblance of male and female GID patients to participants of the same biological sex (Van Goozen et al., 2002; Haraldsen et al., 2005; Sommer et al., 2008). Our study indicates, to the contrary, that the lack of change in performance of GID males receiving cross-sex hormone treatment might be seen as evidence for their not resembling their biological sex, since we found that C males guessed more when being retested, an adjustment in strategy which benefited their performance.

In conclusion, we found that Norwegian males and females did not differ in guessing strategy at the first presentation of a novel test which calls for numerical problem solving. At the third measurement, however, males showed an increase in confidence compared with females. We propose that this finding indicates that men feel most confident when in a testing situation that is out of the ordinary, but not completely new to them. Women seem to be immune to this effect, and stick to a relatively conservative answering strategy. We advise examiners, as well as researchers and testing psychologists, to take this sex difference in guessing strategy (and the change herein) into account when calculating scores based on standardized multiple choice tests, especially when it contains arithmetic subtests, and when interpreting these scores. This could be particularly important when retesting the participant.

Surprisingly, male GID patients undergoing cross-sex hormone treatment did not adjust their guessing strategy at all; after 12 months of cross-sex hormone treatment, the male GID patients guessed significantly less frequently than the C males. This is an important finding because it indicates that even though it has been shown that cross-sex hormone treatment does not result in a change in cognitive performance, it might still have an impact on other psychological traits that indirectly affect performance or an adjustment herein, and have been shown to differ between men and women, such as risk taking behavior and feelings of self-confidence. In fact, a rare study that focused both on psychological traits and cognitive performance showed that the trait 'anger proneness' changed as an effect of cross-sex hormone treatment (Van Goozen et al., 1995). Pinpointing which traits are subject to change as a result of hormonal treatment would be useful for psychologists, who can then prepare the patients for these changes as well as guide the patients through them. We hope that future research will further elucidate the potential influence of cross-sex hormone treatment on personality traits.

Acknowledgments

The study was supported by the South-Eastern Norway Regional Health Authority, the Norwegian Research Council, and the University of Oslo. The authors thank Marijtje van Duijn of the University of Groningen for statistical assistance, Jan van Bebber, Anne-Kristin Solbakk, Torbjørn Elvsåshagen, Mitzi Paap and Anja Martine Jansen for helpful feedback and discussions, and all participants.

References

American Psychiatric Association, 2000. Diagnostic and Statistical Manual of Mental Disorders: DSM-IV-TR. APA, Washington, DC.

Ben-Shakhar, G., Sinai, Y., 1991. Gender differences in multiple-choice tests: the role of differential guessing tendencies. Journal of Educational Measurement 28, 23–35.

Brandell, G., Larsson, S., Nyström, P., Palbom, A., Staberg, E.-M., Sundqvist, C., 2005. Kön och matematik. (Reprints in Mathematical Sciences, 2005:20). Lund, Sweden: Centre for Mathematical Sciences, Lund University.

Cem Demirci, S.F.W., 2008. Congenital adrenal hyperplasia. Dermatologic Therapy 21, 340–353.

Christiansen, K., Knussmann, R., 1987. Sex hormones and cognitive functioning in men. Neuropsychobiology 18, 27–36.

Doolittle, A.E., Cleary, T.A., 1987. Gender-based differential item performance in mathematics achievement items. Journal of Educational Measurement 24, 157–166.

Ekenstierna, M., 2004. Multiple comparison procedures based on marginal p-values. Uppsala University, Uppsala. http://www.math.uu.se/research/pub/Ekenstierna.pdf.

Ekstrom, R.B., French, J.W., Harman, H.H., Dermen, D., 1976. Kit of factor-referenced cognitive tests. Educational Testing Service, Princeton, NJ.

Embretson, S.E., Reise, S., 2000. Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Publishers, Mahwah, N.J.

Eriksson, K., Lindholm, T., 2007. Making gender matter: the role of gender-based expectancies and gender identification on women's and men's math performance in Sweden. Scandinavian Journal of Psychology 48, 329–338.

Geschwind, N., Galaburda, A.M., 1985. Cerebral lateralization. Biological mechanisms, associations, and pathology: I. A hypothesis and a program for research. Archives of Neurology 42, 428–459.

Gierl, M.J., Bisanz, J., Bisanz, G.L., Boughton, K.A., 2003. Identifying content and cognitive skills that produce gender differences in mathematics: a demonstration of the multidimensionality-based DIF analysis paradigm. Journal of Educational Measurement 40, 281–306.

Gouchie, C., Kimura, D., 1990. The relationship between testosterone levels and cognitive ability patterns. Psychoendocrinology 16, 323–334.

Hampson, E., Kimura, D., 1988. Reciprocal effects of hormonal fluctuations on human motor and perceptual spatial skills. Behavioral Neuroscience 3, 456–459.

Haraldsen, I.R., Opjordsmoen, S., Egeland, T., Finset, A., 2003. Sex-sensitive cognitive performance in untreated patients with early onset gender identity disorder. Psychoneuroendocrinology 28, 906–915.

Haraldsen, I.R., Egeland, T., Haug, E., Finset, A., Opjordsmoen, S., 2005. Cross-sex hormone treatment does not change sex-sensitive cognitive performance in gender identity disorder patients. Psychiatry Research 137, 161–174.

Hudson, L., (1986). Item-level analysis of sex differences in mathematics achievement test performance. Dissertation Abstracts International, 47, 850-B.

Huynh, H., Feldt, L.S., 1976. Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. Journal of Educational and Behavioral Statistics 1, 69–82.

Hyde, J.S., 2005. The gender similarities hypothesis. American Psychologist 60, 581–592.

Inzlicht, M., Ben-Zeev, T., 2000. A threatening intellectual environment: why females are susceptible to experiencing problem-solving deficits in the presence of males. Psychological Science 11, 365–371.

James, T.W., Kimura, D., 1997. Sex differences in remembering the location objects in an array: location-shifts versus location-exchanges. Evolution and Human Behavior 18, 155–163.

Keller, J., Dauenheimer, D., 2003. Stereotype threat in the classroom: dejection mediates the disrupting threat effect on women's math performance. Personality and Social Psychology Bulletin 29, 371–381.

Kimball, M.M., 1989. A new perspective on women's math achievement. Psychological Bulletin 105, 198–214.

Linn, M.C., Petersen, A.C., 1985. Emergence and characterization of sex differences in spatial ability: a meta-analysis. Child Development 56, 1479–1498.

Liu, O.L., Wilson, M., Paek, I., 2008. A multidimensional Rasch analysis of gender differences in PISA mathematics. Journal of Applied Measurement 9, 18–35.

Maccoby, E.E., Jacklin, C.N., 1974. The Psychology of Sex Differences. Stanford University Press, Stanford, Calif.

Malouf, M.A., Migeon, C.J., Carson, K.A., Petrucci, L., Wisniewski, A.B., 2006. Cognitive outcome in adult women affected by congenital adrenal hyperplasia due to 21-hydroxylase deficiency. Hormone Research 65, 142–150.

McGillicuddy-De Lisi, A.V., De Lisi, R., 2002. Biology, Society, and Behavior: The Development of Sex Differences in Cognition. Ablex Publishing, Westport, Conn.

McKeever, W.F., 1995. Hormone and hemisphericity hypotheses regarding cognitive sex differences: possible future explanatory power, but current empirical chaos. Learning and Individual Differences 7, 323–340.

Meurling, A.W., Tonning-Olsson, I., Levander, S., 2000. Sex differences in strategy and performance on computerized neuropsychological tests as related to gender identity and age at puberty. Scandinavian Journal of Psychology 41, 81–90.

Puts, D.A., McDaniel, M.A., Jordan, C.L., Breedlove, S.M., 2008. Spatial ability and prenatal androgens: meta-analyses of congenital adrenal hyperplasia and digit ratio (2D:4D) studies. Archives of Sexual Behavior 37, 100–111.

Resnick, S.M., Berenbaum, S.A., Gottesman, I.I., Bouchard, T.J., 1986. Early hormonal influences on cognitive functioning in congenital adrenal hyperplasia. Developmental Psychology 22, 191–198.

Shaffer, J.P., 1995. Multiple hypothesis testing. Annual Review of Psychology 46, 561–584.

Sherwin, B., 1994. Estrogenic effects on memory in women. Annals of the New York Academy of Sciences 743, 213–230.

Silverman, I., Phillips, K., 1993. Effects of estrogen changes over the menstrual cycle on spatial performance. Ethology and Sociobiology 14, 250–270.

Slabbekoorn, D., van Goozen, S.H., Megens, J., Gooren, L.J., Cohen-Kettenis, P.T., 1999. Activating effects of cross-sex hormones on cognitive functioning: a study of short-term and long-term hormone effects in transsexuals. Psychoneuroendocrinology 24, 423–447.

Sommer, I.E.C., Cohen-Kettenis, P.T., van Raalten, T., vd Veer, A.J., Ramsey, L.E., Gooren, L.J.G., Kahn, R.S., Ramsey, N.F., 2008. Effects of cross-sex hormones on cerebral activation during language and mental rotation: an fMRI study in transsexuals. European Neuropsychopharmacology 18, 215–221.

SPSS for Windows, Rel. 15.0.1.1, 2007. Chicago: SPSS Inc.

Stevens, J.P., 2002. Applied Multivariate Statistics for the Social Sciences. Lawrence Erlbaum Associates, Mahwah, NJ.

Torres, A., Gomez-Gil, E., Vidal, A., Puig, O., Boget, T., Salamero, M., 2006. Gender differences in cognitive functions and influence of sex hormones. Actas Españolas de Psiquiatría 34, 408–415.

Van Goozen, S.H., Cohen-Kettenis, P.T., Gooren, L.J., Frijda, N.H., Van de Poll, N.E., 1995. Gender differences in behaviour: activating effects of cross-sex hormones. Psychoneuroendocrinology 20, 343–363.

Van Goozen, S.H., Slabbekoorn, D., Gooren, L.J., Sanders, G., Cohen-Kettenis, P.T., 2002. Organizing and activating effects of sex hormones in homosexual transsexuals. Behavioral Neuroscience 116, 982–988.

Voyer, D., Voyer, S., Bryden, M., 1995. Magnitude of sex differences in spatial abilities: a meta-analysis and consideration of critical variables. Psychological Bulletin 117, 250–270.

Wester, A., Henriksson, W., 2000. The interaction between item format and gender differences in mathematics performance based on TIMSS data. Studies in Educational Evaluation 26, 79–90.

Williams, C., Meck, W., 1990. The organizational effects of gonadal steroids on sexual dimorphic spatial ability. Psychoneuroendocrinology 16, 155–176.